Es un extracto del libro: Hernández Pérez, A., 1999. La búsqueda y recuperación de información en Internet. En: M. Caridad Sebastián (Editor), La sociedad de la información: política, tecnología e industria de los contenidos. Centro de Estudios Ramón Areces, Madrid, pp. 213-242.

Motores o robots de búsqueda.

Aunque en la actualidad directorios y motores de búsqueda tienden a confluir, existen todavía diferencias significativas entre ellos. La diferencia fundamental es que los directorios son realizados mediante una categorización realizada por personas mientras que los motores de búsqueda utilizan sistemas totalmente automatizados para construir sus bases de datos. Para explicar el funcionamiento de los motores o máquinas de búsqueda de Internet distinguiremos los distintos elementos que intervienen en el proceso:

- El robot. Que se encarga de la construcción de la base de datos. El modo en que los robots (*spiders, crawlers,* etc.) construyen las bases de datos.
- El sistema de indexación automática.
- El lenguaje de interrogación. La potencialidad del servicio, respecto al hardware -qué tipo y cuantas máquinas atienden a tantos usuarios- y respecto al software, si permite utilizar operadores booleanos, búsqueda por campos (por rangos de fecha, por idioma de los documentos, por título), etc.
- El interfaz de usuario. Para la de introducción de datos (si permite búsqueda simple y búsqueda avanzada) y para la presentación de resultados (si limita el número de resultados, si los ordena, si ofrece sólo el título o una descripción más amplia del resultado de lo que encuentra), etc.

8.1. Construcción de las base de datos: el robot.

Para construir la base de datos, el motor de búsqueda utiliza un conjunto de dos programas que son los que conforman el robot: un localizador de recursos y un recolector.

El *Localizador* cuenta con una lista de servidores que debe visitar. Se dirige al primer servidor y/o sitio web que tiene asignado y hace un recorrido por él buscando en anchura y/o en profundidad. O sea, se dedica a localizar dentro de la página de ese servidor los enlaces hipertextos que encuentra. Cuando encuentra un enlace el localizador comprueba si tiene o no tiene ya el enlace y, si no lo tiene, lo almacena en su base de datos. El localizador puede optar entre seguir el enlace hacia otra página o continuar examinando la misma página. Habitualmente los localizadores tienen límites para no exceder un número de niveles de directorios desde la página de comienzo. A medida que va realizando su trabajo el *Localizador* envía la información recogida al *Recolector.*

El *Recolector* hace una comprobación en su base de datos de direcciones con el fin de averiguar si ya tiene recogida la dirección que le ha entregado el localizador. Si la tiene, comprueba el tamaño y la fecha para detectar si ha habido cambios en el documento localizado. Si ha habido cambios o si no tiene la dirección, captura el documento y se lo entrega al analizador. Si por el contrario detecta que en el documento no ha habido cambios desde la última vez que capturó la página, pasa a ocuparse de la siguiente dirección. A medida que va finalizando las comprobaciones va enviando sus resultados al *Analizador*.

8.2. El sistema de indexación automática: el analizador – indexador.

El *Analizador o indexador* es el programa encargado de extraer palabras del contenido de acuerdo a un conjunto de filtros. Estos filtros son los que deciden si se debe indexar toda la página o sólo las primeras 250 palabras, si indexará todas las etiquetas de un documento HTML o sólo el texto contenido en las etiquetas META, TITLE (<TITLE>Título</TITLE>), META, Hx (<H1>Encabezamiento</H1>), etc. y varias decenas de posibilidades más.

El resultado de la tarea del Analizador-indexador es un fichero inverso que será gestionado por una base de datos sobre la que buscarán los usuarios. El proceso es continuo y recursivo de tal forma que los robots van acumulando cada vez más y más direcciones de recursos para analizar. Altavista, uno de los motores más grandes en tamaño, capturaba en enero de 1999 unas 10 millones de páginas por día.

Los criterios para generar la base de datos varían de un motor a otro por lo que las bases de datos resultantes son muy dispares entre sí. Algunos de los factores que influyen en la construcción de estas bases de datos son:

- Los sitios web que deben visitar con más frecuencia, en función de estos sitios web los robots tomarán unas rutas u otras y por tanto descubrirán unos recursos antes que otros.
- El tiempo con que el robot visita vuelve a visitar las páginas ya analizadas. Mientras más tiempo tarda más riesgo de que muchas páginas ya no existan o que la información que tenemos en nuestra base de datos no esté actualizada.
- El tiempo que tarda un robot en indexar una nueva página, lo que varía dependiendo de si la página ha sido descubierta por el robot automáticamente en su fase de rastreo o de si ha sido enviada por el administrador de un sitio web.

- De la profundidad con que analiza los sitios web. Si analiza todas las páginas de un sitio o sólo una muestra.
- De la profundidad con que analiza las páginas web. Si analiza todo el texto y todas las etiquetas de una página o sólo las primeras 250 ó 300 palabras o sólo ciertas etiquetas.

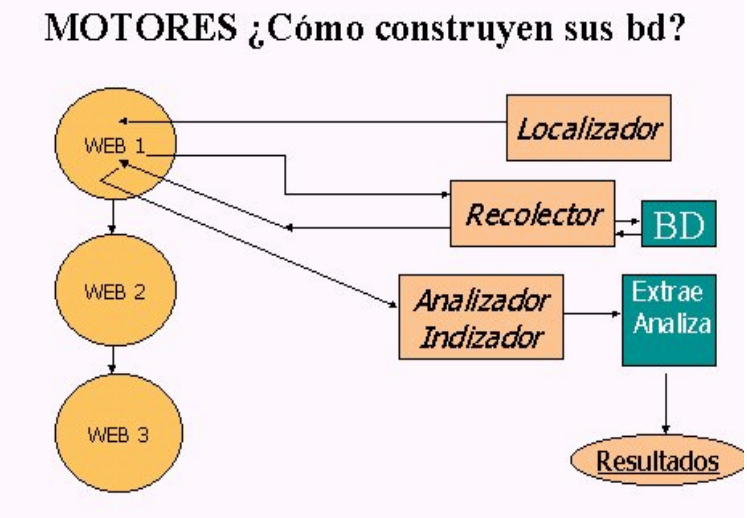

Fig. 1: *Construcción de las bases de datos en los motores de búsqueda.*

8.3. El lenguaje de interrogación de la base de datos.

El siguiente aspecto a estudiar en los motores es su comportamiento ante una consulta. Y aquí hay que distinguir dos aspectos: uno relacionado con el hardware y otro relacionado con el software, con las posibilidades de búsqueda que ofrece el sistema. El relacionado con el hardware tiene que ver con la capacidad de los equipos en los que se alberga la/s bases de datos. Si tienen procesadores rápidos, si tienen sistemas capaces de procesar múltiples consultas concurrentes sin largos tiempos de espera, etc. Este aspecto influirá lógicamente en el grado de aceptación del servicio por parte del usuario.

El aspecto que más nos interesa ahora es el de las posibilidades de búsqueda que ofrecen estos lenguajes de recuperación y estas posibilidades van desde las muy limitadas en unos robots a las muy potentes en otros. Lo que sí debe tener siempre en cuenta un usuario a la hora de buscar en estos motores de búsqueda es que el lenguaje de recuperación que utilizan es diferente en cada uno de ellos. Algunos motores utilizan el operador de unión "or" por defecto cuando tecleamos dos palabras mientras que otros utilizan por defecto el operador de intersección "and". Hay motores que hacen el truncamiento de forma automática mientras que otros no permiten los truncamientos y muchas otras diferencias más. Dado que

las posibilidades de recuperación cambian de unos a otros es siempre recomendable acceder a los archivos de ayuda de cada uno de ellos.

No obstante, algunas de las características o posibilidades que permiten estos lenguajes de recuperación de información tienen que ver con:

- La búsqueda de palabras simples: es la forma de búsqueda más habitual que actualmente reciben los motores de búsqueda. El usuario introducen una palabra simple. Por ejemplo, "cine", "periodismo", etc.
- La búsqueda de referencias compuestas: otra de las formas de búsqueda más populares. El usuario introduce un concepto compuesto por dos palabras. Por ejemplo, "cine negro", "periodismo electrónico". Los sistemas responden de muy diversas formas a este tipo de búsquedas. Algunos combinan los dos términos de tal forma que pueden salir documentos que hablen de periodismo y que contengan la palabra electrónico, pero que no hablen de periodismo electrónico. Para la búsqueda de este tipo de frases lo más recomendable es emplear siempre comillas.
- La búsqueda con operadores booleanos. Una de las características más extendidas. Utiliza la lógica booleana para extraer documentos de la base de datos. Los operadores más comunes son los de unión (OR), los de intersección (AND) y los de exclusión (NOT). Algunos motores utilizan el operador de unión como operador por defecto cuando realizamos una búsqueda que contiene dos términos y en la que el usuario no introduce los términos entrecomillas. Por ejemplo, buscar por *periodismo electrónico* sin utilizar las comillas dará un resultado muy distinto al de buscarlo utilizando comillas. En el primer caso, se extraerán todos los documentos que contienen la palabra *periodismo,* todos los que contengan la palabra *electrónico* y todos los que contengan ambas palabras, estén o no juntas. O sea, muchos documentos que no tendrán que ver con el tema. En el segundo caso, *"periodismo electrónico",* el sistema sólo extraerá aquellos documentos en los que ambas palabras aparezcan juntas. Para evitar el ruido que causa utilizar el operador de unión (OR), muchos motores utilizan ahora el operador de intersección (AND) con lo que se reduce el ruido pero pueden salir documentos que contengan los términos periodismo y electrónico pero sin ninguna relación entre ellos.
- La búsqueda con operadores relacionales. Son los operadores de comparación, muy útiles y menos extendido de lo normal en los motores de búsqueda. Sirven, por ejemplo, para acotar el resultado de una búsqueda sólo a los documentos publicados después de 1998.
- La búsqueda de referencias adyacentes. Un operador no muy común en los motores actuales. Le pedimos a la base de datos que encuentre documentos que contengan dos palabras pero que se

encuentren en el mismo párrafo o en la misma frase. Por ejemplo, "historia" NEAR "Altavista" extrae documentos de la base de datos que tienen que ver con la historia de Altavista. No basta con que aparezcan en el documento las palabras "historia" y "Altavista" sino que además deben aparecer en el mismo párrafo.

- La búsqueda cualificada, la búsqueda por campos. Incomprensiblemente muy poco implementada y muy poco utilizada por los usuarios en aquellos motores, como Altavista, en que se permite. Sirve para limitar la búsqueda de uno o varios términos a uno o varios campos. Se puede acotar para que busque las palabras que representan los temas que nos interesan sólo en el campo o en el contenido de la etiqueta título (TITLE) o en el contenido de la etiqueta META. O por ejemplo acotar una búsqueda a los servidores existentes en España, en Altavista utilizando el campo *domain*, (domain:es) o limitándola a uno o a todos los servidores de la Universidad Carlos III de Madrid, etc.
- La búsqueda con truncamiento. Los truncamientos suelen ser siempre por la derecha y significa que el motor busca en la base de datos por la raíz de la palabra que buscamos. Por ejemplo, que al buscar por periodis* encuentre todo lo concerniente a periodismo, periodista, periodistas, etc. Algunos motores utilizan esta opción sin ni siquiera informar al usuario de tal forma que aunque solicitemos búsquedas por periodismo el sistema realiza un truncamiento automático.
- La búsqueda distinguiendo entre mayúsculas y minúsculas. Algunos motores no distinguen entre mayúsculas y minúsculas pero otros sí de tal forma que los resultados a una búsqueda serán distintos si el texto a buscar se introduce todo en mayúscula, todo en minúscula o la primera letra con mayúscula y el resto con minúscula. Como norma casi general se puede decir que si la búsqueda se introduce toda en minúsculas el sistema busca todas las palabras estén como estén escritas.
- La búsqueda teniendo en cuenta un fichero de palabras vacías. Algunos motores no indexan palabras que no tienen ningún significado como las preposiciones, artículos, etc. lo que mejora los tiempos de indexación y de búsqueda en la base de datos.
- Combinación de búsquedas. Muy poco implementados. La posibilidad de realizar búsquedas en diversas etapas y combinar a posteriori el resultado de dos o más búsquedas.
- Búsquedas con sistemas de retroalimentación de relevancia. Sistemas que permiten que tras ver el resultado de una búsqueda el usuario pueda seleccionar uno o más documentos para que el motor de búsqueda trate de encontrar, por patrones de frecuencias de palabras, documentos parecidos a los seleccionados por el usuario.
- Búsquedas por pesos. Búsquedas con más de un término en las que se permite al usuario definir cuales de los términos son más importantes en la búsqueda.

Los problemas de la recuperación de información tienen que ver con las posibilidades de búsqueda de los distintos sistemas pero, sobre todo, son problemas lingüísticos. Piense por ejemplo que al buscar por "periodismo electrónico" estamos perdiendo todos los documentos que puedan hablar de "periodismo digital" o de periodistas digitales o los documentos que hablen de nuevas tendencias periodísticas en Internet, etc. Algunos de los problemas más importantes[1] tienen que ver con la sinonimia, con la homonimia, con las falsas combinaciones de términos, con las ambigüedades sintácticas, con la polisemia, con la dificultad para generalizar, etc.

8.4. El interfaz de usuario: la introducción y la visualización de los documentos.

La interfaz de usuario tiene que ver fundamentalmente con dos aspectos: la forma en que el sistema permite al usuario introducir sus búsquedas y la forma en que presenta los resultados de esas búsquedas. Respecto al primero los sistemas mantienen interfaces muy homogénea basadas en formularios muy simples en los que al usuario se le presenta una sola casilla en la que debe escribir las palabras por las que desea buscar. Algunos otros presentan formularios algo más complejos, con dos o más recuadros para escribir, botones, recuadros y listas de opciones (para elegir, por ejemplo, el idioma de los documentos que desea recuperar o para limitar los años sobre los que desea realizar la búsqueda, etc.) pero en general la mayoría utiliza formularios simples.

El segundo aspecto, el de la presentación de resultados, es mucho más interesante puesto que cuando una base de datos devuelve un resultado de más de 100 documentos es muy importante el orden en que se presentan esos documentos porque el usuario tenderá a visualizar sólo aquellos que son presentados en primer lugar. La importancia es tal que algunos motores de búsqueda han comenzado a cobrar por ofrecer un sitio web en un puesto de los primeros lugares de la pantalla de resultados cuando el usuario realiza búsquedas por palabras concreta[2]. Por ejemplo, que una empresa de venta de muebles aparezca en los cinco primeros lugares cuando alguien introduce en una búsqueda la palabra "mueble" o la palabra "mesa", etc.

La cuestión clave respecto a los interfaces es pues el orden en que son mostrados esos documentos. Las fórmulas que emplean los distintos motores para componer su algoritmo de relevancia,

---

[1] Un excelente trabajo sobre el tema se puede encontrar en: Frederick W. Lancaster. *El control de vocabulario en la recuperación de información*. Valencia: Universitat de Valencia, 1995

[2] Véases por ejemplo el caso de GoTo en http://www.goto.com

que es lo que determina el orden en que se muestran los documentos, se ha convertido casi en secreto industrial. Prácticamente todos utilizan fórmulas que tienen que ver con la frecuencia de aparición de un término y los algoritmos clásicos de las teorías de la recuperación de información, como el de la frecuencia inversa del término de Salton[3], de Blair[4] y de otros autores.

Algunos de los factores que influyen en el orden en que se presentan los resultados de una búsqueda son:

- Si los términos que se buscan son parte del contenido de las etiquetas META que describen el documento.
- Si los términos que se buscan son parte del contenido de las etiquetas TITLE, Hx, STRONG y otras etiquetas que destacan el valor de un término o una frase en un documento basado en lenguaje de marcas.
- La frecuencia de uno o más términos en el documento.
- La frecuencia de uno o más términos en el total de la base de datos.
- La frecuencia de coocurrencia de dos términos en los documentos.

Respecto a la visualización de la pantalla de resultados también hay cierto consenso entre los diferentes motores. Como norma general ofrecen en una pantalla sólo los 10 ó 20 primeros resultados para no ralentizar la descarga de una página de resultados excesivamente grande y a continuación en otras páginas otro número similar. Prácticamente todos muestran el título del documento, una breve descripción y algunos incluso el tamaño en Kb del documento y la fecha de la última modificación del mismo.

Metabuscadores y multibuscadores.

La existencia de tantos motores de búsqueda y la diferencia de resultados que presentan cada uno de ellos a una misma búsqueda a causa de la distintas técnicas para construir las respectivas bases de datos han dado lugar a la generación de dos nuevos sistemas de búsqueda: los metabuscadores y los multibuscadores.

[3] Gerald Salton. *Introduction to modern information retrieval*. New York: Mac-Graw Hill, 1983. También, Gerald Salton. *Automatic text processing*. Reading: Addinson-Wesley, 1989

[4] David C. Blair. *Language and representation in information retrieval*. Amsterdam: Elsevier, 1990

Los metabuscadores permiten que un usuario pueda realizar desde una única pantalla una búsqueda y lanzarla al mismo tiempo contra múltiples motores de búsqueda (Lycos, Infoseek, Excite, Altavista, etc.). Los metabuscadores utilizan las bases de datos de otros y ellos se ocupan de ordenar los resultados que van llegando eliminando duplicados, permitiendo elegir diversos formatos de visualización de resultados, etc. La utilidad de los metabuscadores reside pues en realizar búsquedas simultáneas en diferentes motores de búsqueda.

Los multibuscadores tienen un funcionamiento más simple puesto que, a diferencia de los metabuscadores, su única utilidad es recoger en una misma página los códigos necesarios para realizar una búsqueda en diferentes motores distintos. Aunque en la actualidad también tienden a converger, la diferencia inicial estriba en que en los metabuscadores basta introducir una vez el término o los términos de búsqueda para que busque en todas las bases de datos mientras que en los multibuscadores es necesario introducir los términos una vez por cada buscador. Pero la gran diferencia es que los metabuscadores ofrecen un valor añadido, búsqueda simultánea, eliminación de duplicados, elección de formatos de visualización, etc., que por ahora los multibuscadores no ofrecen.

Al igual que ocurre con los directorios la construcción de motores de búsqueda genéricos como Altavista requieren una altísima inversión tecnológica. Por un lado tienen que albergar gran cantidad de documentos y por otro, tienen que conseguir que la búsqueda se efectúe con rapidez. Por esta razón, la tendencia actual es construir *motores de búsqueda especializados*. En estos casos el motor sólo se dirige a los sitios que le marca su creador, normalmente restringidos a un área geográfica o temática o incluso una mezcla de ambos. En este tipo de motores el robot no tiene un comportamiento "libre" sino que está limitado a ciertos servidores.

Tecnologías *"push"*, canales y agentes de búsqueda y recuperación de información.

Existen dos métodos para obtener información en la red: uno, el que hemos visto hasta ahora mediante directorios o motores, es el que emplean desde siempre los usuarios/as, y consiste en ir a buscar la información, modelo *pull*. El otro es el modelo *push*, en el que los usuarios/as no tienen que ir a buscar la información, sino que son las fuentes de información las que en teoría, mediante técnicas de filtrado de información, les envían sólo lo que realmente les interesa, de acuerdo a un perfil de interés marcado por ellos mismos. Este modelo push es el que emplean los canales de información de algunos sitios web. Con este tipo de servicios es posible obtener información actualizada acerca de los temas que nos interesen en un momento dado sin necesidad de tener que ir a buscarla.

El funcionamiento de la tecnología *push* está basado fundamentalmente en el envío de información al cliente, por parte de un servidor que actúa como canal, a través de la red Internet. Para que un cliente reciba información de un canal es necesario que se suscriba previamente a éste. Una vez establecida la suscripción al canal éste envía información al usuario en el momento que éste indique y una vez descargada la información leerla desde su propio ordenador sin necesidad de estar conectado a Internet.

Los ***canales*** pues no son más que servidores Web, o sea, un conjunto de páginas que muestran información al usuario en formato HTML o XML. El contenido de la información es pues el mismo que se encuentra en la sede Web de la institución a la que el usuario se conecta sólo que es enviada al usuario sin que éste acuda a buscarla. La información que se envía por los canales puede presentarse de cuatro formas:

- En forma de mensaje de correo electrónico, enviando al usuario/a el texto completo por correo electrónico.
- En forma de página HTML o XML, enviando al usuario/a por correo electrónico el URL que les lleva a la información.
- En forma de componente del escritorio.
- En forma de protector de pantalla, en donde la información enviada se va mostrando cuando se activa el protector.
- O mediante la ejecución y lectura de un programa *ad hoc* como Pointcast[5].

Este nuevo modelo de obtención de información, que recuerda a la clásica difusión selectiva de la información utilizada por los grandes distribuidores de información desde los años 70, supone una ayuda de gran valor puesto que, una vez seleccionadas las fuentes de información y los temas que interesan, el usuario puede ahorrar mucho tiempo en la obtención de información. Las críticas a estos sistemas activos de recuperación de información provienen de la escasa madurez de los sistemas de filtros de información por lo que los usuarios se quejan de recibir excesiva información que no les interesa y que echan de menos información publicada que no han recibido.

Dentro de estos modelos *push* en los que la información le llega al usuario en vez de ser éste el que vaya a buscarla existen los llamados agentes de búsqueda. ***Los agentes*** *pueden ser considerados*

[5] Para más información véase http://www.pointcast.com

como asistentes personales de software con autoridad delegada por parte de sus usuarios[6]. Patti Maes, una de las investigadoras principales del Media Lab (en el MIT) afirma que es posible pasar de la metáfora de la manipulación directa, que requiere que el usuario inicie una tarea y controle todos los acontecimientos a la metáfora de la gestión indirecta, en la que el usuario se ve inmerso en un proceso cooperativo, con un programa de ordenador llamado asistente personal inteligente que podría, por ejemplo, filtrar las noticias, gestionar el correo electrónico, arreglar las citas, seleccionar libros y discos y otras formas de entretenimiento[29].

Maes ha identificado dos problemas con estos agentes:

1. Un problema de competencia: cómo adquiere el agente el conocimiento para decidir cuando, con qué y cómo ayudar al usuario?.
2. Un problema de confianza: cómo se asegura que los usuarios se sienten cómodos delegando tareas en el agente?.

De acuerdo con Maes, estos problemas se pueden resolver con una aproximación al aprendizaje de las máquinas en el que el agente aprende acerca de los hábitos de su usuario a través de las interacciones. Así, un agente adquiriría su competencia de la forma siguiente: observando e imitando al usuario, recibiendo *feedback* positivo y negativo del usuario, recibiendo instrucciones explícitas del usuario y preguntando a otros agentes para que le den consejos.

En los últimos años han ido apareciendo en Internet algunos agentes inteligentes[7], con capacidad limitada de aprendizaje y autónomos para la toma de decisiones, que no sólo se dedican a buscar y recuperar información sino también a organizarla y filtrarla. La mayoría de estos agentes tienen que ver con la adquisición de productos (discos y libros fundamentalmente) y con la búsqueda y filtrado de información procedente de revistas profesionales y diarios de actualidad.

[6] Fah-Chun Cheong. *Internet agents: spiders, wanderers, brokers and bots*. Indianapolis: New Riders, 1996

[7] Un ejemplo de estos programas en el que está implicada la propia Maes se puede encontrar en Firefly: http://www.firefly.net